Direct Whole-Genome Sequencing of Cutaneous Strains of *Haemophilus ducreyi*

Michael Marks,[1] Maria Fookes,[1] Josef Wagner, Rosanna Ghinai, Oliver Sokana, Yaw-Adu Sarkodie, Anthony W. Solomon, David C.W. Mabey, Nicholas R. Thomson

Haemophilus ducreyi, which causes chancroid, has emerged as a cause of pediatric skin disease. Isolation of *H. ducreyi* in low-income settings is challenging, limiting phylogenetic investigation. Next-generation sequencing demonstrates that cutaneous strains arise from class I and II *H. ducreyi* clades and that class II may represent a distinct subspecies.

Since 2000, the global prevalence of chancroid, caused by *Haemophilus ducreyi*, has declined (*1*). *H. ducreyi* is an emerging cause of cutaneous ulcers in tropical countries (*1–4*). Cutaneous lesions of *H. ducreyi* are difficult to distinguish from other common causes of ulcerative skin disease, such as yaws (*3,4*), which presents problems in diagnosing yaws and has resulted in the World Health Organization recommending molecular testing of yaws-like lesions (*5*).

Culturing *H. ducreyi* is challenging. PCR is usually used for diagnosis (*6*). Culture requirements limit sequencing and phylogenetic analyses. Traditional phylogenies divide genital strains of *H. ducreyi* into class I and II clades. Most studies suggest that cutaneous strains of *H. ducreyi* have diversified from within the class I clade (*7,8*), and a recent study reported cutaneous strains that appear to arise from class II strains (*9*). These studies have been limited by the number and geographic spread of samples included.

Next-generation sequencing enables whole-genome sequencing from clinical samples without prior culture, bypassing the culture requirements of *H. ducreyi* and enabling more detailed phylogenetic analysis. We performed next-generation sequencing on samples collected in

Author affiliations: London School of Hygiene & Tropical Medicine, London, UK (M. Marks, R. Ghinai, A.W. Solomon, D.C.W. Mabey, N.R. Thomson); Hospital for Tropical Diseases, London (M. Marks, A.W. Solomon, D.C.W. Mabey); Wellcome Trust Sanger Centre, Cambridge, UK (M. Fookes, J. Wagner, N.R. Thomson); Solomon Islands Ministry of Health and Medical Services, Honiara, Solomon Islands (O. Sokana); Kwame Nkrumah University of Science and Technology, Kumasi, Ghana (Y.-A. Sarkodie)

previous surveys conducted in the Solomon Islands (in 2013) and Ghana (in 2014) (*2,4*). In both surveys, skin swab specimens had been collected from persons with chronic ulcerative lesions believed, at the time, to be clinically consistent with yaws. DNA was prepared for the current study from samples with residual material from those original surveys. The London School of Hygiene & Tropical Medicine, Solomon Islands National Health Research, and Kwame Nkrumah University of Science and Technology ethics committees approved these studies.

The Study

We tested 72 samples from 63 persons (Figure 1). Twenty-five persons (27 samples) had been recruited in Ghana and 38 persons (45 samples) in the Solomon Islands. Median age of participants in the original studies was 9 years (interquartile range 7–11 years); 36 (57.1%) were male. In the original studies, 24 samples had tested positive for *H. ducreyi* using a 16S rRNA-targeted PCR (*2,4*): 15 from the Solomon Islands and 9 from Ghana.

In Ghana, samples were collected directly onto dry Dacron swabs. In the Solomon Islands, swab exudate was placed into transport medium (AssayAssure; Sierra Molecular, Incline Village, NV, USA) or onto an FTA Elute Card (Thermo-Fisher Scientific, Waltham, MA, USA). Samples were frozen at –20°C and shipped to the Centers for Disease Control and Prevention (Atlanta, GA, USA) on dry ice for the original laboratory analyses, which included real-time PCR for *Trepomema pallidum* subspecies *pertenue* (*7*) and a real-time 16S rRNA-targeted PCR for *H. ducreyi* (*2,4*). After testing, samples were shipped on dry ice to the London School of Hygiene & Tropical Medicine (London, UK) and frozen at –20°C before analysis.
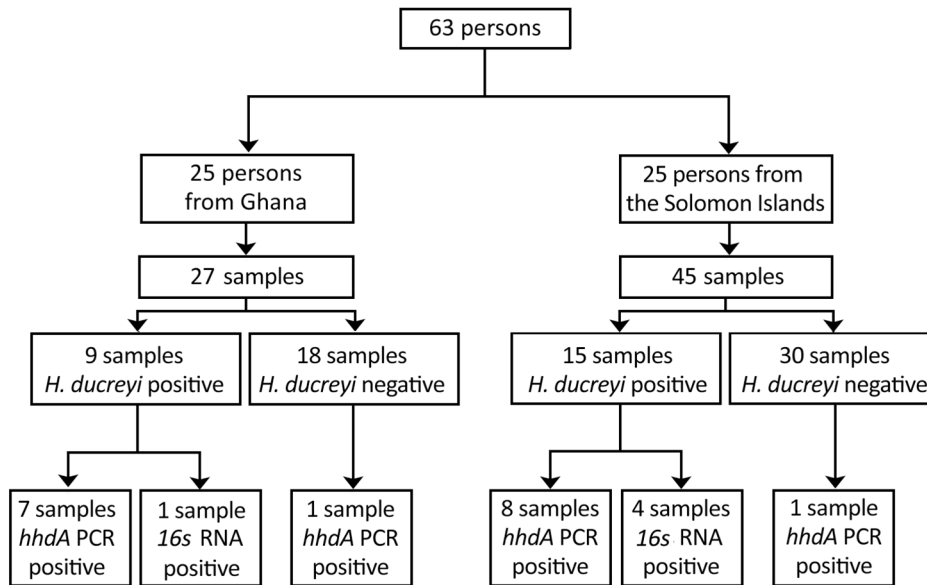
We extracted DNA from residual sample material using QIAamp Mini kits (QIAGEN, Hilden, Germany) (online Technical Appendix 1, https://wwwnc.cdc.gov/EID/article/24/4/17-1726-Techapp1.pdf). We screened DNA using a quantitative PCR (qPCR) targeting the *hhdA* gene and 16S rRNA gene sequencing for *H. ducreyi* (*6,10*). From samples that tested positive, we selected those with genomic DNA concentration ≥10 copies/μL for direct (non–culture-based) sequencing.

Genomic DNA was fragmented to an average size of 150 bp and subjected to DNA library creation using

---

Figure. Flowchart of whole-genome sequencing of *Haemophilus ducreyi*. Samples were originally collected in 2 studies conducted in Ghana (2014) and the Solomon Islands (2013) (*2,4*). Results of the *H. ducreyi* PCR conducted in the original studies and of the 2 *H. ducreyi* PCRs performed in this study are shown.

established Illumina paired-end protocols (*11*). We amplified adaptor-ligated libraries and indexed them by PCR. We used a portion of each library to create an equimolar pool and hybridized each pool to custom-made SureSelect RNA baits (Agilent Technologies, Santa Clara, CA, USA; baits based on published sequences of *H. ducreyi* [*12*]) (online Technical Appendix 1). Targets were captured and amplified in accordance with manufacturer's recommendations. We subjected enriched libraries to standard 75-bp end sequencing (HiSeq 2000; Illumina, San Diego, CA, USA). Samples' public accession numbers are listed in online Technical Appendix 2 Table 1 (https://wwwnc.cdc.gov/EID/article/24/4/17-1726-Techapp2.xlsx). We used whole-genome sequence data to estimate phylogenies for *H. ducreyi* (online Technical Appendix 1), including publicly available *H. ducreyi* genomes alongside those obtained in this study.

We identified *H. ducreyi* in 17 samples by *hhdA*-targeted qPCR and in 5 additional samples using an assay targeting the rRNA gene. From these 22 positive samples, we obtained 21 (95.5%) complete genomes from 13 persons from the Solomon Islands and 8 from Ghana. Mean coverage of *H. ducreyi* genomes was 91% (online Technical Appendix 2 Table 1). We found no evidence of sequence heterozygosity that would indicate any participant was infected with multiple distinct strains of *H. ducreyi*.

H. ducreyi sequences fell into both previously defined *H. ducreyi* clades: class I and class II (Figure 2). To estimate genetic distance between strains, we determined the number of single-nucleotide polymorphisms (SNPs) in pairwise whole-genome comparisons. The average distance between class I and class II sequences was 21,238 SNPs, compared with a maximum pairwise distance of

641 SNPs between class I sequences. We detected 4 major recombination blocks within class I genomes. These regions included the *dsrA*, *tad,* and *flp* loci, associated with serum resistance, tight adhesion, and production of fimbriae, respectively, functions important in micro-colony formation and potentially associated with virulence (online Technical Appendix 1 Figure 1; online Technical Appendix 2 Table 2) (*13*). The other regions of likely recombination were related to integrated prophage elements, implying *H. ducreyi* has an actively exchanging bacteriophage repertoire in its genome (online Technical Appendix 1 Figure 1). These prophage elements included the region coding for the ctdABC genes, which have been associated with virulence (*14*). The class I prophage elements were absent from class II genomes but intermittently present in class I genomes (online Technical Appendix 1 Figure 2). The presence or absence of the ctdABC coding region was not associated with cutaneous or genital ulcer disease. Another recombination region spanned the *hhdA* specific qPCR primer binding site. Samples with high sequence variation in this region tested negative for *H. ducreyi* by qPCR but gave high numbers of reads by 16S rRNA gene sequencing.

Conclusions

We obtained whole-genome sequences of *H. ducreyi* without prior culture. Most earlier studies have suggested that cutaneous strains emerged by diversification from within the class I clade (*7,8*), although 1 study found, in keeping with our findings, cutaneous strains emerging from class II (*9*). We found genital and cutaneous strains are represented in all lineages of the expanded phylogenetic tree (*7*). We found considerable genetic variation between class I and
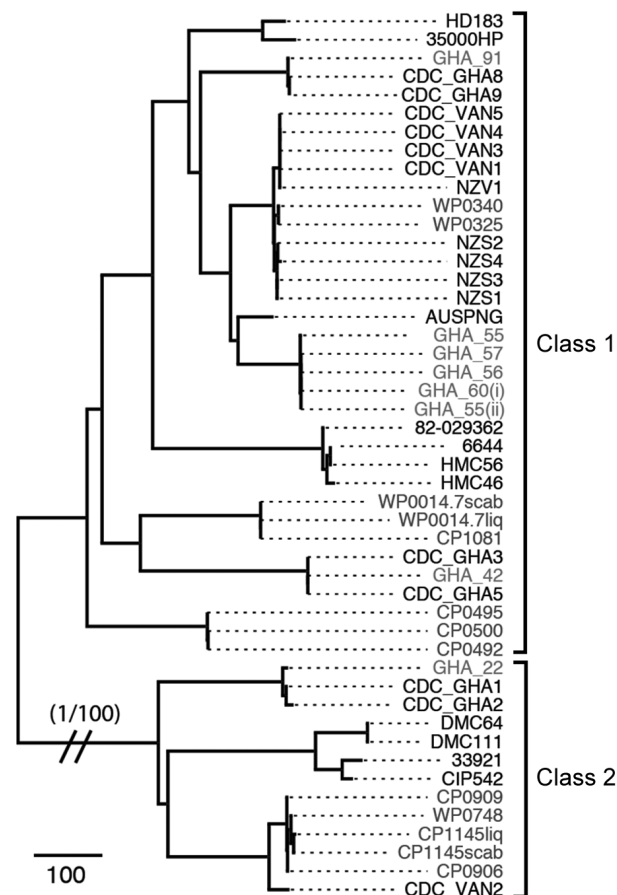
Figure 2. Phylogenetic tree of *Haemophilus ducreyi* genome sequences inferred from mapping using the *H. ducreyi* 35000HP strain as reference and after removing high-density single-nucleotide polymorphisms regions with Gubbins (3). Included are published genomes (black text), Ghanaian strains (gray text, GHA designations), and Solomon Islands strains (gray text, CP/WP designations). Sequences from cutaneous ulcers in Ghana and the Solomon Islands were found within both previously described clades of *H. ducreyi* class I and class II. Scale bar indicates nucleotide substitutions per site. An expanded version of this figure providing complete phylogeny details, including countries of origin, years, ulcer types, and genome region designations, is provided in online Technical Appendix 1 Figure 1 (https://wwwnc.cdc.gov/EID/article/24/4/17-1726-Techapp1.pdf).

class II *H. ducreyi* sequences. Together with existing 16S rRNA data and multilocus sequence typing data (*12*) these findings suggest class II strains might represent a discrete subspecies of *H. ducreyi*.

We identified 2 samples that had been negative in the original studies but were found to contain *H. ducreyi* DNA in the current study. Repeated freeze–thaw cycles and limited residual DNA volumes might have contributed to our lack of detection of *H. ducreyi* DNA in 4 samples that tested positive in the original studies (Figure 1). Five samples that returned a weak signal by *hhdA* qPCR contained class II clade

H. ducreyi genomes. The failure of qPCR to detect *H. ducreyi* in these samples was most likely explained by variation in the sequence of the *hhdA* pPCR primer binding sites (*13*) between class I and II genomes (online Technical Appendix 1 Figure 1), demonstrating our limited understanding of the diversity of these pathogens.

Culture for *H. ducreyi* is not practical in the low-income settings where cutaneous strains of this organism are endemic. Next-generation sequencing circumvents this problem by enabling whole-genome sequencing direct from clinical samples. This approach considerably strengthens our ability to sequence *H. ducreyi* and broaden knowledge of this emerging pathogen.

About the Author

Dr. Marks is an assistant professor at the London School of Hygiene & Tropical Medicine. His primary research interests are the control of neglected tropical diseases, particularly yaws and scabies.

References

1. González-Beiras C, Marks M, Chen CY, Roberts S, Mitjà O. Epidemiology of *Haemophilus ducreyi* infections. Emerg Infect Dis. 2016;22:1–8. http://dx.doi.org/10.3201/eid2201.150425
2. Ghinai R, El-Duah P, Chi K-H, Pillay A, Solomon AW, Bailey RL, et al. A cross-sectional study of 'yaws' in districts of Ghana which have previously undertaken azithromycin mass drug administration for trachoma control. PLoS Negl Trop Dis. 2015;9:e0003496. http://dx.doi.org/10.1371/journal.pntd.0003496
3. Mitjà O, Lukehart SA, Pokowas G, Moses P, Kapa A, Godornes C, et al. *Haemophilus ducreyi* as a cause of skin ulcers in children from a yaws-endemic area of Papua New Guinea: a prospective cohort study. Lancet Glob Health. 2014;2:e235–41. http://dx.doi.org/10.1016/S2214-109X(14)70019-1
4. Marks M, Chi K-H, Vahi V, Pillay A, Sokana O, Pavluck A, et al. *Haemophilus ducreyi* associated with skin ulcers among children, Solomon Islands. Emerg Infect Dis. 2014;20:1705–7. http://dx.doi.org/10.3201/eid2010.140573
5. Marks M, Mitjà O, Vestergaard LS, Pillay A, Knauf S, Chen C-Y, et al. Challenges and key research questions for yaws eradication. Lancet Infect Dis. 2015;15:1220–5. http://dx.doi.org/10.1016/S1473-3099(15)00136-X
6. Chen C-Y, Ballard RC. The molecular diagnosis of sexually transmitted genital ulcer disease. In: MacKenzie CR, Henrich, editors. Diagnosis of sexually transmitted diseases—methods and protocols. New York: Springer-Verlag; 2012. p. 103–12.
7. Gangaiah D, Webb KM, Humphreys TL, Fortney KR, Toh E, Tai A, et al. *Haemophilus ducreyi* cutaneous ulcer strains are nearly identical to Class I genital ylcer strains. PLoS Negl Trop Dis. 2015;9:e0003918. http://dx.doi.org/10.1371/journal.pntd.0003918
8. Gangaiah D, Marinov GK, Roberts SA, Robson J, Spinola SM. Draft whole-genome sequence of *Haemophilus ducreyi* strain AUSPNG1, isolated from a cutaneous ulcer of a child from

Papua New Guinea. Genome Announc. 2016;4:e01661–15. http://dx.doi.org/10.1128/genomeA.01661-15

9. Gangaiah D, Spinola SM. *Haemophilus ducreyi* cutaneous ulcer strains diverged from both Class I and Class II genital ulcer strains: implications for epidemiological studies. PLoS Negl Trop Dis. 2016;10:e0005259. http://dx.doi.org/10.1371/journal.pntd.0005259

10. Marks M, Vahi V, Sokana O, Puiahi E, Pavluck A, Zhang Z, et al. Mapping the epidemiology of yaws in the Solomon Islands: a cluster randomized survey. Am J Trop Med Hyg. 2015;92:129–33. http://dx.doi.org/10.4269/ajtmh.14-0438

11. Orle KA, Gates CA, Martin DH, Body BA, Weiss JB. Simultaneous PCR detection of *Haemophilus ducreyi, Treponema pallidum*, and herpes simplex virus types 1 and 2 from genital ulcers. J Clin Microbiol. 1996;34:49–54.

12. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina

13. Ricotta EE, Wang N, Cutler R, Lawrence JG, Humphreys TL. Rapid divergence of two classes of *Haemophilus ducreyi*. J Bacteriol. 2011;193:2941–7. http://dx.doi.org/10.1128/JB.01400-10

14. Janowicz DM, Cooney SA, Walsh J, Baker B, Katz BP, Fortney KR, et al. Expression of the Flp proteins by *Haemophilus ducreyi* is necessary for virulence in human volunteers. BMC Microbiol. 2011;11:208. http://dx.doi.org/10.1186/1471-2180-11-208

sequencing system. Nat Methods. 2008;5:1005–10. http://dx.doi.org/10.1038/nmeth.1270

Address for correspondence: Michael Marks, Clinical Research Department, London School of Hygie ne & Tropical Medicine, Keppel Street, London WC1E 7HT, UK; email: michael.marks@lshtm.ac.uk; Nicholas Thomson, Wellcome Trust Sanger Centre, Hinxton CB10 1SA, UK; email: nrt@sanger.ac.uk

Direct Whole-Genome Sequencing of Cutaneous Strains of *Haemophilus ducreyi*

Technical Appendix 1

DNA Extraction Methods Used

Dry Dacron swab tips from Ghana were suspended in 1 mL of lysis buffer (10 mM Tris, pH 8.0; 0.1M EDTA, pH 8.0; 0.5% sodium dodecyl sulphate), vortexed vigorously for 5 min, and the supernatants divided into 3 tubes. The Qiagen (Hilden, Germany) buccal swab spin protocol (which includes a proteinase K reaction step) was then followed. The same procedure was applied to Solomon Islands samples (transported in AssayAssure transport buffer, Sierra Molecular, Incline Village, USA); lesion crusts were suspended in the liquid buffer and pelleted, and DNA extractions were performed on both the liquid phase (as above) and pellet. DNA extraction of the lesion pellet was performed as follows: pellets were resuspended in 500 mL of lysis buffer, treated with proteinase K, and transferred to a Lysing Matrix E tube (MP Biomedicals, Santa Ana, CA, USA) prefilled with various sized glass beads. 200 mL of AL buffer was added to this. Filled tubes were homogenized for 40 sec at speed 6,000 rpm (FastPrep Homogenizer, MP Biomedicals). The tubes were then centrifuged for 10 min at 13,200 rpm and the Qiagen DNA extraction protocol followed as above. The DNA was then extracted according to manufacturers' instructions. Each DNA sample was eluted in 150 μL AE buffer.

SureSelect Probe Design

SureSelect probe design was based on str. 35000HP (EMBL accession no. AE017143). Agilent 120 bp RNA baits were designed to cover the pan-genome consensus sequence to a depth of 3×. RNA baits were screened against the pan-genome consensus sequence for redundancy (see http://earray.chem.agilent.com/suredesign). All captured DNA was validated by sequencing and mapping to an appropriate reference genome.

Bio-informatics Pipeline

Reads were mapped (*1*) using SMALT to the reference genomes *of H. ducreyi* strain 35000HP (GenBank accession no. NC_002940), respectively. Whole-genome single nucleotide polymorphisms (SNPs) were identified using SamTools (*2*). Gubbins was used to identify recombination blocks (defined by having high SNP density) using the whole-genome SNP data of each sample as previously described (*3*). SNPs in these recombination blocks were excluded from phylogenetic analysis because they do not represent the underlying phylogeny of the host (Technical Appendix 1 Figure 2; Technical Appendix 2 Table 2, https://wwwnc.cdc.gov/EID/article/24/4/17-1726-Techapp2.xlsx). The remaining chromosomal SNP alleles from each isolate were concatenated to generate a multiple alignment of all SNPs, per isolate. Maximum likelihood phylogenetic trees were estimated from these remaining SNPs using RAxML with a general time reversible site model with gamma correction for among-site variation and with 100 bootstrap replicates (*4*). De novo genome assemblies were performed as previously described (*5*) or using SPAdes (*6*). Contigs were automatically annotated using in-house pipelines. Genes of interest were identified and curated by hand in pairwise comparisons with the appropriate reference genome using the Artemis Comparison Tool or Artemis (*7*). Basic mapping and assembly statistics are listed in re listed in Technical Appendix 2 Table 1 (https://wwwnc.cdc.gov/EID/article/24/4/17-1726-Techapp2.xlsx) , including the in silico simulated reassemblies of the 2 reference sequences to determine the achievable reassembly size from using Illumina 75-bp paired-end reads and the standard assembly protocols used in this study on the simulated reference sequence data.
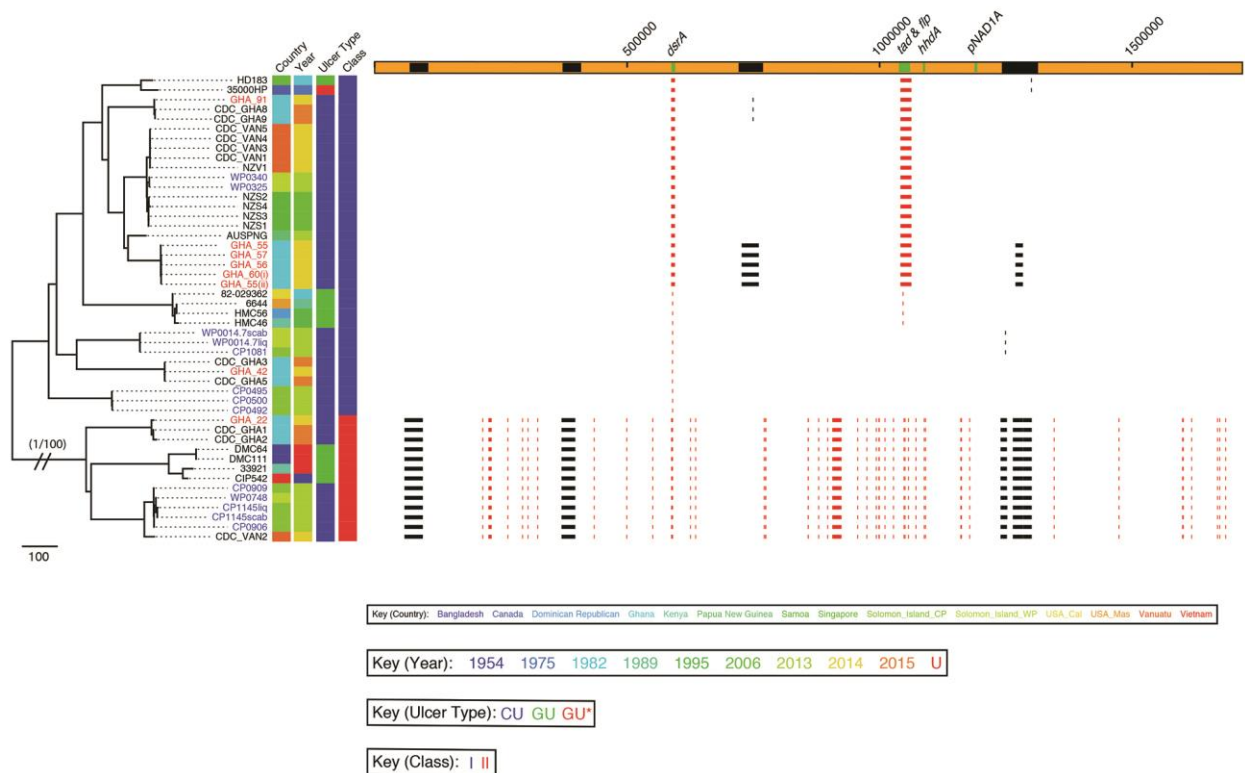
References

1. Sanger Institute. SMALT [cited 2016 Dec 20]. http://www.sanger.ac.uk/science/tools/smalt-0

2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. PubMed

3. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15. PubMed

4. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90. PubMed

5. Bronowski C, Fookes MC, Gilderthorp R, Ashelford KE, Harris SR, Phiri A, et al. Genomic characterisation of invasive non-typhoidal *Salmonella enterica* subspecies *enterica* serovar Bovismorbificans isolates from Malawi. PLoS Negl Trop Dis. 2013;7:e2557. PubMed

6. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77. PubMed

7. Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, et al. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics. 2008;24:2672–6. PubMed
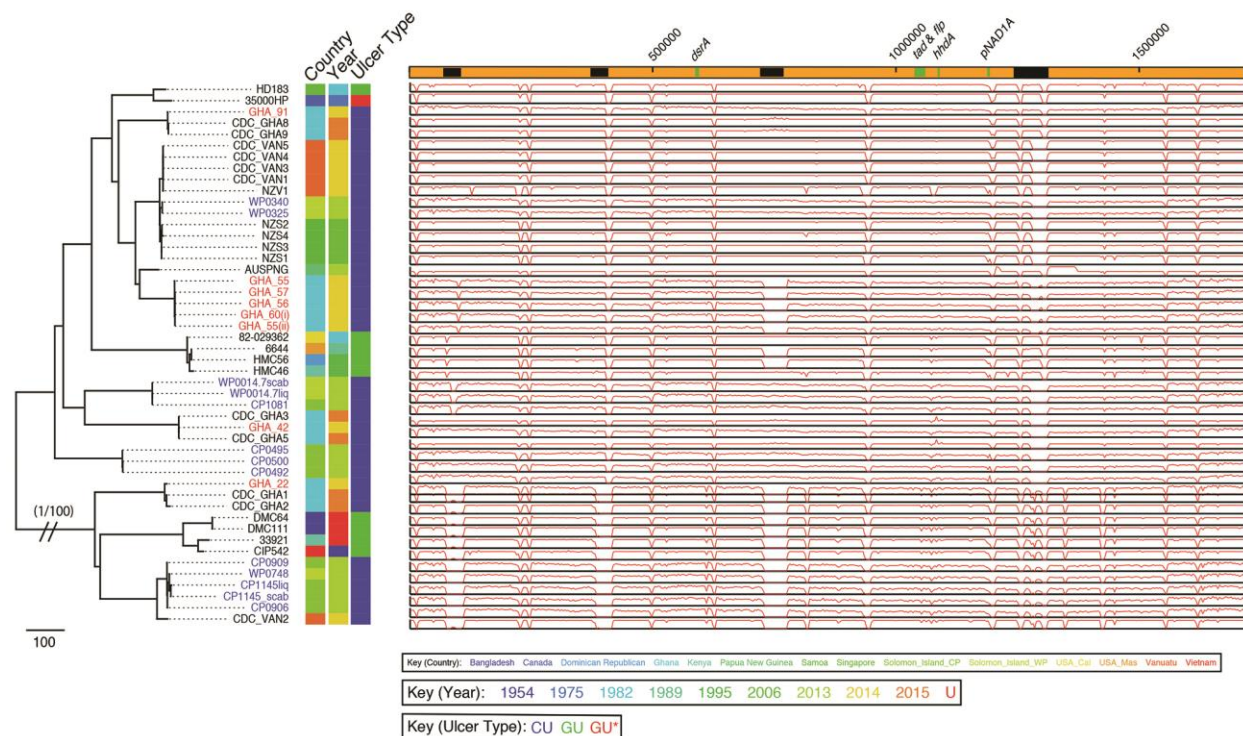
Technical Appendix 1 Figure 1. Phylogenetic tree of *Haemophilus ducreyi* genomes. Phylogenetic tree of *H. ducreyi* sequences inferred from mapping using the *H. ducreyi* 35000HP strain as reference and after removing high-density single-nucleotide polymorphisms regions with Gubbins (*3*). These regions are shown in the main frame below the genome with red for those events shared by >1 strains and blue for

those unique to a single strain. Included are published genomes (black labels), Ghanaian strains (red labels), and Solomon Islands strains (blue labels). Over the genome (orange bar), prophage-related regions are black and other relevant regions are green and labeled. Coordinates are in bps with respect to the reference. Sequences from cutaneous ulcers in Ghana and the Solomon Islands were found within both previously described clades of *H. ducreyi* Class I and Class II. Scale bar indicates nucleotide substitutions per site.

Technical Appendix 1 Figure 2. *Haemophilus ducreyi* genome sequence coverage. Coverage of genomes mapped to the reference 35000HP strain. Over the genome, phage locations are in black. Class I prophage elements were absent from Class II genomes but intermittently present in Class I genomes. Scale bar indicates nucleotide substitutions per site.